Expression of Plastid-Targeted Polypeptides in Plants

This invention relates to methods and means for the expression of plastid-targeted polypeptides in plants.

Plastids are membrane-bound organelles within plant cells which have a variety of cellular functions. Examples of plastids include chloroplasts, proplastids, chromoplasts, etioplasts and leucoplastids, such as amyloplasts and proteinoplasts.

Although some plastid proteins are encoded by plastid DNA and synthesised within the plastid, most plastid proteins are encoded by the nuclear genome and synthesized in the cytosol as precursors. These precursors contain an amino-terminal transit peptide that is both necessary and sufficient to direct the transport of the precursor from the cytosol, across the outer and inner envelope membranes, into the plastid stroma, where the transit peptide is cleaved off to generate the mature protein (Keegstra, K. & Cline, K. *Plant Cell* 11 557-570 (1999)). In the chloroplast, for example, a hetero-oligomeric molecular machine known as the Tic/Toc translocon complex (Soll, J. *Curr. Opi Plant Biol.* 5, 529-535 (2002)), which is located in the chloroplast envelope membranes, mediates the specific recognition and translocation of precursor proteins into the chloroplast.

The present inventors have recognised that certain plastid-localised proteins in plants are not, in fact, targeted directly to the plastid from the cytosol but are instead directed to the endoplasmic reticulum and become glycosylated before entering the plastid stroma. This finding has significant utility in the expression of recombinant polypeptides in plants.

One aspect of the invention provides a method of producing a recombinant polypeptide comprising;

expressing in a plant cell a nucleic acid encoding a fusion polypeptide which comprises said recombinant polypeptide, an ER signal sequence and one or more ER-plastid targeting sequences.

The expressed fusion polypeptide may subsequently be cleaved to produce said recombinant polypeptide.

The ER signal sequence and one or more ER-plastid targeting sequences are preferably heterologous to the recombinant polypeptide. The ER signal sequence and one or more ER-plastid targeting sequences may be from the same or different sources.

The ER signal sequence directs the localisation of the polypeptide from the cytosol to the ER. A suitable ER signal sequence may comprise at least 20 amino acids, at least 22 amino acids or at least 24 amino acids. The ER signal sequence is preferably a plant ER signal sequence, for example a plant ER signal sequence from the N terminal of an ER-processed plastid polypeptide. Examples of ER-processed plastid polypeptides from chloroplasts are listed in Table 1.

Examples of suitable ER signal sequence include;
MKIMMMIKLCFFSMSLICIAPADA,
MAASHGNAIFVLLLCTLFLPSLAC, and;
MAARIGIFSVFVAVLLSISAFSSA.

Other examples of ER signal sequences are described in Emanuelsson et al *J. Mol. Biol.* 300, 1005-1016 (2000).

ER-plastid targeting sequences direct the transit of polypeptides within the plant cell from the microsomes (i.e. the ER or Golgi) to a plastid, which may, for example, be a proplastid, chromoplast, etioplast, leucoplastid (e.g. amyloplast or proteinoplast) or chloroplast. In some preferred embodiments, the ER-plastid targeting sequence is an ER-chloroplast targeting sequence which directs the transit of a polypeptide to the chloroplast.

A suitable ER-plastid targeting sequence may comprise a sequence of at least 10 contiguous amino acids, more preferably 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120 or more contiguous amino acids from an ER-processed plastid polypeptide or an allele, variant or derivative thereof, in particular from the N or C terminal of an ER-processed plastid polypeptide or an allele, variant or derivative thereof. A targeting sequence from an ER-processed polypeptide from a particular plastid may be used to target polypeptide to that plastid. In some preferred embodiments, the full-length sequence of an ER-processed plastid polypeptide or an allele, variant or derivative thereof may be employed i.e. the one or more ER-plastid targeting sequences are comprised within an ER processed plastid polypeptide. Examples of ER-processed plastid polypeptides found in the chloroplast are listed in Table 1. ER-processed plastid polypeptides from other plastids, for example proplastids, chromoplasts, etioplasts, or leucoplastids, may be readily identified using standard techniques, as described herein.

One, two, three or more ER-plastid targeting sequences may be employed within a fusion polypeptide as described herein.

In some embodiments, an ER-plastid targeting sequence may comprise or consist of a 12 to 15 amino acid sequence from the C terminal of an ER-processed plastid polypeptide. Such a sequence may be hydrophilic and, in some preferred embodiments, may comprise 2, 3, 4 or more contiguous basic residues, in particular lysine and/or arginine residues. For example, an ER-plastid targeting sequence may be comprise or consist of the amino acid sequence KKETGNKKKKPN, RFWGKKKRRSSP or TGKKKKKTYLP. Other suitable sequences may be obtained from the C terminal region (i.e. the C terminal 20-30 amino acids) of a sequence from the list shown in Table 1.

In some embodiments, the one or more ER-plastid targeting sequence may comprise or consist of residues 25 to 114 and/or residues 224 to 285 of a CAH1 polypeptide, for example *A. thaliana* CAH1. In some

preferred embodiments, the fusion protein may further comprise an ER signal sequence comprising or consisting of residues 1 to 24 of CAH1 as described above. Thus, a fusion polypeptide may comprise, in an N to C direction, residues 1 to 114 of CAH1, a sequence encoding a recombinant polypeptide, and residues 224 to 285 of CAH1. In some particularly preferred embodiments, the fusion polypeptide may comprise the full-length CAH1 sequence.
The recombinant polypeptide may be upstream (i.e. towards the N terminal) or downstream (i.e. towards the C terminal) of the one or more ER-plastid targeting sequences within the fusion polypeptide, or may be located between two or more ER-plastid targeting sequences.

For example, in some embodiments, a recombinant polypeptide may be joined directly or indirectly to the N terminal or C terminal of an ER-processed plastid polypeptide within the fusion polypeptide, or may be located within the ER-processed plastid polypeptide sequence (i.e. surrounded by sequence from the ER-processed plastid polypeptide).

Recombinant polypeptide may be generated from the fusion polypeptide by any convenient means. Typically, proteolytic cleavage of the fusion polypeptide using one or more endoproteases may be employed. Suitable endoproteases may include site-specific endoproteases, such as rennin, factor Xa and thrombin, or other endoproteases known in the art.

In some embodiments, an endoprotease may be present within the plastid, either as an endogenous plant polypeptide, such as SPP, (Richter et al J. Biol. Chem. (2002) 277: 43888-43894), DEG (Itzhaki et al J. Biol. Chem. (1998) 273: 7094-7098) or FTSH, or as a recombinant polypeptide expressed from a heterologous nucleic acid. The expressed fusion polypeptide may thus undergo *in situ* proteolysis to produce the recombinant polypeptide within the plastid.

To facilitate cleavage by endoproteases, the recombinant polypeptide sequence may be linked to heterologous sequences within the fusion polypeptide, such as the ER signal sequence and ER-plastid targeting sequences, by cleavable linkers. Suitable linker sequences are well known in the art and may include, for example substrate sequences for thrombin, rennin, and factor X. Other suitable linker sequences are described in Richter et al J. Biol. Chem. (2002) 277: 43888-43894.

After cleavage of the fusion polypeptide to produce the recombinant polypeptide, the recombinant polypeptide may be isolated and/or purified from the plastid. Plastids may be isolated from the plant cell in a preliminary purification, prior to purification of the recombinant polypeptide from the isolated plastids. Alternatively, recombinant polypeptide may be isolated directly from the plant cells.

In other embodiments, the fusion polypeptide may be isolated and/or purified from the plastid prior to the generation of the recombinant polypeptide. For example, the fusion polypeptide may be isolated and treated with endoproteases to liberate the recombinant polypeptide.

Expressed polypeptide may be extracted, isolated and/or purified from plants or plant material by any convenient method. For example, the plant material may be homogenised, solvent extracted and subjected to chromatographic separation methods such as HPLC and column chromatography, for example using a silica column. In some embodiments, the expressed polypeptide is glycosylated and glycosylation-specific purification methods may be employed, for example using a column containing immobilised lectin or glycosyl-specific antibodies.

In some preferred embodiments, a recombinant polypeptide may be produced in accordance with the invention by expressing in a plant cell a nucleic acid encoding a fusion polypeptide which comprises

said recombinant polypeptide linked to an ER-processed plastid polypeptide.

The recombinant polypeptide may subsequently be cleaved from the ER-processed plastid polypeptide.

The recombinant polypeptide or the fusion polypeptide may be isolated and/or purified from the plastid following said expression.

As described above, the ER processed plastid polypeptide may be positioned downstream (i.e. towards the C terminal) or more preferably upstream (i.e. towards the N terminal) of the recombinant polypeptide, or may be located within the ER-processed plastid polypeptide sequence (i.e. surrounded by sequence from the ER-processed plastid polypeptide).

Preferably, the fusion polypeptide comprises an N terminal ER signal sequence. In embodiments in which the ER-processed plastid polypeptide is upstream of the recombinant polypeptide, the ER signal sequence may be comprised within the ER-processed plastid polypeptide sequence.

An ER processed plastid polypeptide is a polypeptide located in the plastid which is post-translationally targeted to the plastid via the ER. Suitable ER processed plastid polypeptides may be identified by standard *in silico* analysis and data mining techniques. For example, ER processed chloroplast polypeptides may be identified from sequences obtained by chloroplast proteome initiatives (Friso, G et al (2004) Plant Cell (in press), T. Kleffmann, et al (2004) Current Biology (in press)). Examples of ER processed chloroplast polypeptides from these databases, which contain an ER signal peptide but lack a C-terminal H/KDEL ER-retention signal, are listed in Table 1. Gene ID's are based on the Arabidopsis Genome Initiative (Nature (2000) 408(6814):796-815).

ER processed plastid polypeptides may comprise an N-terminal ER signal sequence as identified by targetP predictions. They may further comprise a hydrophilic C- or N-terminal, for example comprising 2 or more basic residues, in particular lysines and/or arginine residues.

In some embodiments, an ER processed plastid polypeptide may comprise one or more glycosylation sites, preferably N-glycosylation sites. These sites may be glycosylated when the polypeptide is expressed in plant cells.

Suitable ER processed plastid polypeptides include Arabidopsis CAH1 (U73462), Rice CAH1 (CAD40654), Arabidosis ribophorin 1 and other sequences which are listed in Table 1.

Whilst a wild-type ER processed plastid polypeptide is preferred in the fusion polypeptides described herein, an ER processed plastid polypeptide which is a fragment, mutant, derivative, variant or allele of such a wild type sequence may also be used

Suitable fragments, mutants, derivatives, variants and alleles of ER processed plastid polypeptides retain the signals required for targeting to the plastid via the ER. A mutant, variant or derivative may have one or more of addition, insertion, deletion or substitution of one or more amino acids in the polypeptide sequence. Such alterations may be caused by one or more of addition, insertion, deletion or substitution of one or more nucleotides in the encoding nucleic acid.

A polypeptide which is an amino acid sequence variant, allele, derivative or mutant of an ER processed plastid polypeptide such as CAH1, for example Arabidopsis CAH1 (U73462), or a sequence listed in Table 1, may comprise an amino acid sequence which shares greater than about 30% sequence identity with the wild-type polypeptide sequence, greater than about 35%, greater than about 40%, greater than about 45%, greater than about 55%, greater than

about 65%, greater than about 70%, greater than about 80%, greater than about 90% or greater than about 95%. The sequence may share greater than about 30% similarity with the wild-type ER processed plastid polypeptide sequence, greater than about 40% similarity, greater than about 50% similarity, greater than about 60% similarity, greater than about 70% similarity, greater than about 80% similarity or greater than about 90% similarity.

Sequence similarity and identity are commonly defined with reference to the algorithm GAP (Genetics Computer Group, Madison, WI). GAP uses the Needleman and Wunsch algorithm to align two complete sequences that maximizes the number of matches and minimizes the number of gaps. Generally, default parameters are used, with a gap creation penalty = 12 and gap extension penalty = 4. Use of GAP may be preferred but other algorithms may be used, e.g. BLAST (which uses the method of Altschul *et al.* (1990) *J. Mol. Biol.* 215: 405-410), FASTA (which uses the method of Pearson and Lipman (1988) *PNAS USA* 85: 2444-2448), or the Smith-Waterman algorithm (Smith and Waterman (1981) *J. Mol Biol. 147*: 195-197), or the TBLASTN program, of Altschul et al. (1990) supra, generally employing default parameters. In particular, the psi-Blast algorithm (Nucl. Acids Res. (1997) 25 3389-3402) may be used. Sequence identity and similarity may also be determined using Genomequest™ software (Gene-IT, Worcester MA USA).

Sequence comparisons are preferably made over the full-length of the relevant sequence described herein.

Similarity allows for "conservative variation", i.e. substitution of one hydrophobic residue such as isoleucine, valine, leucine or methionine for another, or the substitution of one polar residue for another, such as arginine for lysine, glutamic for aspartic acid, or glutamine for asparagine.

The recombinant polypeptide which is expressed using the methods described herein may be any polypeptide of interest. The present

methods are particularly suitable for the expression of glycosylated polypeptides. Suitable polypeptides may include vaccines (for example, vaccines against hepatitis B virus envelope protein, human cytomegalovirus glycoprotein B or Norwalk virus capsid protein), antibodies or antibody fragments, pharmaceutical proteins such as signal peptides, protein hormones, structural proteins such as collagen, blood proteins such as serum albumin, enzymes such as secreted alkaline phosphatase, industrial enzymes and enzymes that produce a secondary or new metabolite/chemical compound in the plastid. Other examples of recombinant polypeptides are described in Trends in Plant Science (2001) 6 5 219-226 and Ma et al Nature Reviews Genetics 4, 794 -805 (2003).

In some preferred embodiments, the recombinant polypeptide may comprise one or more N-glycosylation sites (for example Asn-x-Thr/Ser sites) and/or O-glycosylation sites. Targeting to the plastid via the microsomes allows the glycosylation of such sites. Methods as described herein are therefore especially suitable for the production of glycosylated recombinant polypeptides. The presence or amount of glycosylation, for example by a xylose- or fucose-containing glycan, may be determined following production of the recombinant polypeptide in the plant. Glycosylation may be determined by any convenient method. For example, the polypeptide may be contacted with an antibody specific for a glycosyl epitope, such as $\beta(1,2)$-xylose or $\alpha(1,3)$-fucose.

Methods of the invention allow the recombinant polypeptide to pass through the ER and the Golgi system, enabling N- and O-glycosylation and maturation of the glycosylation pattern. The glycosylation pattern may be a plant glycosylation pattern, for example comprising $\beta(1,2)$-xylose and/or $\alpha(1,3)$-fucose residues. This is exemplified herein by the presence, in the glycosylated CAH1 protein described below, of fucose, which is added in the Golgi. In other embodiments, the glycosylation pattern may be a

mammalian glycosylation pattern, for example comprising $\alpha(1,6)$-fucose residues.

A recombinant polypeptide expressed as described herein may thus comprise N- and/or O linked glycosyl residues.

Another aspect of the invention provides a nucleic acid construct comprising a nucleotide sequence which encodes an ER signal sequence and one or more ER-plastid targeting sequences, the nucleotide sequence further comprising one or more restriction endonuclease sites (i.e. a cloning site), which are preferably suitable for insertion of a nucleotide coding sequence capable of expressing a recombinant (i.e. a heterologous) polypeptide fused to said ER signal and plastid targeting sequences.

ER signal sequences and plastid targeting sequences are described above.

The nucleic acid construct may further comprise a nucleotide coding sequence encoding a recombinant polypeptide for expression as part of said fusion polypeptide, said coding sequence being inserted in the cloning site. The invention encompasses an isolated nucleic acid comprising a nucleotide sequence which encodes a fusion protein in which a recombinant polypeptide is fused to an ER signal sequence and one or more ER-plastid targeting sequences.

In some embodiments, the nucleotide sequence encoding the ER-plastid targeting sequences, and preferably also the ER signal sequence, may be comprised within a nucleotide sequence encoding an ER processed plastid polypeptide. According to such embodiments, a nucleic acid construct may comprise a nucleotide sequence which encodes an ER processed plastid polypeptide and one or more restriction endonuclease sites for insertion of a nucleotide coding sequence capable of expressing a recombinant polypeptide fused to said ER processed plastid polypeptide.

Suitable ER processed plastid polypeptides are described in more detail above.

The nucleic acid construct may further comprise a nucleotide sequence encoding one or more cleavable linkers which allow the liberation of the recombinant polypeptide from the fusion polypeptide after expression. For example, the recombinant polypeptide may be fused to the ER signal sequence and ER-plastid targeting sequences by a cleavable linker. Suitable linkers may be cleaved by a site-specific endoprotease such as thrombin, factor Xa or rennin.

The nucleotide sequence encoding the fusion polypeptide may be operably linked to a heterologous regulatory sequence.

The regulatory sequence or element may be plant specific i.e. it may preferentially direct the expression (i.e. transcription) of a nucleic acid within a plant cell relative to other cell types. For example, expression from such a sequence may be reduced or abolished in non-plant cells, such as bacterial or mammalian cells.

The heterologous regulatory sequence may be activated by a heterologous transcription factor, such as GAL4 or T7 polymerase. Nucleic acid encoding the heterologous transcription factor may be operably linked to a plant-specific promoter as described above so that expression of the heterologous transcription factor is plant specific and plant specific expression of the fusion polypeptide by activation of the heterologous regulatory sequence. For example, a GAL4 transcription factor may be expressed using a CaMV35S promoter and may drive expression of a fusion polypeptide coding sequence which is operably linked to the GAL4 promoter. In other embodiments, T7 polymerase may be expressed using a CaMV35S promoter and may drive expression of a coding sequence which is operably linked to a T7 promoter.

The terms "heterologous" and "recombinant" are used to indicate that the sequence of nucleotides in question has been introduced into a nucleic acid construct or a plant cell or an ancestor thereof, using genetic engineering or recombinant means, i.e. by human intervention and is not naturally found in such a construct or cell. A sequence which is heterologous (i.e. exogenous or foreign) to another nucleotide sequence or host cell is not associated with that sequence or cell in nature.

A heterologous plant specific regulatory sequence may be an inducible promoter. Such a promoter may induce expression in response to a stimulus. This allows control of expression, for example, to allow optimal plant growth before fusion polypeptide production is induced.

The term "inducible" as applied to a promoter is well understood by those skilled in the art. In essence, expression under the control of an inducible promoter is "switched on" or increased in response to an applied stimulus (which may be generated within a cell or provided exogenously). The nature of the stimulus varies between promoters. Whatever the level of expression is in the absence of the stimulus, expression from any inducible promoter is increased in the presence of the correct stimulus. The preferable situation is where the level of expression increases in the presence of the relevant stimulus by an amount effective to cause production of polypeptide. Thus an inducible (or "switchable") promoter may be used which causes a basic level of expression in the absence of the stimulus which causes little or no accumulation of polypeptide. Upon application of the stimulus, which may for example, be an increase in environmental stress, expression of polypeptide is increased (or switched on).

Many examples of inducible promoters will be known to those skilled in the art.

Other suitable promoters may include the Cauliflower Mosaic Virus 35S (CaMV 35S) gene promoter that is expressed at a high level in virtually all plant tissues (Benfey et al, (1990) EMBO J 9: 1677-1684); the cauliflower meri 5 promoter that is expressed in the vegetative apical meristem as well as several well localised positions in the plant body, e.g. inner phloem, flower primordia, branching points in root and shoot (Medford, J.I. (1992) *Plant Cell* 4, 1029-1039; Medford *et al*, (1991) *Plant Cell* 3, 359-370) and the *Arabidopsis thaliana* LEAFY promoter that is expressed very early in flower development (Weigel *et al*, (1992) *Cell* 69, 843-859). Other suitable promoters may be tissue specific, for example seed or leaf specific, and/or specifically expressed at different times or developmental stages, for example diurnally active promoters such as the CAH1 promoter.

The construct may further comprise a 5′ untranslated region to control translational initiation efficiency and transcript stability and thereby enhance expression.

Nucleic acid sequences and constructs as described above may be comprised within a vector. Those skilled in the art are well able to construct vectors and design protocols for recombinant gene expression, for example in a microbial or plant cell. Suitable vectors can be chosen or constructed, containing appropriate regulatory sequences, including promoter sequences, terminator fragments, polyadenylation sequences, enhancer sequences, marker genes and other sequences as appropriate. A vector may comprise a selectable marker to facilitate selection of the transgenes under an appropriate promoter. For further details see, for example, *Molecular Cloning: a Laboratory Manual:* 3rd edition, Sambrook & Russell, 2001, Cold Spring Harbor Laboratory Press.

Many known techniques and protocols for manipulation of nucleic acid, for example in preparation of nucleic acid constructs, mutagenesis, sequencing, introduction of DNA into cells and gene expression, and analysis of proteins, are described in detail in

Protocols in Molecular Biology, Second Edition, Ausubel et al. eds. John Wiley & Sons, 1992. Specific procedures and vectors previously used with wide success upon plants are described by Bevan, Nucl. Acids Res. (1984) 12, 8711-8721), and Guerineau and Mullineaux, (1993) Plant transformation and expression vectors. In: Plant Molecular Biology Labfax (Croy RRD ed) Oxford, BIOS Scientific Publishers, pp 121-148.

A method of producing a recombinant polypeptide as described herein may comprise incorporating a nucleic acid encoding a fusion polypeptide which comprises said recombinant polypeptide, an ER signal sequence and one or more ER-plastid targeting sequences and;

expressing said nucleic acid to produce a recombinant polypeptide in a plastid of said cell

When incorporating or introducing a chosen gene construct into a cell, certain considerations must be taken into account, well known to those skilled in the art. The nucleic acid to be inserted should be assembled within a construct or vector which contains effective regulatory elements which will drive transcription. There must be available a method of transporting the constructor vector into the cell. Once the construct is within the cell, integration into the endogenous chromosomal material either will or will not occur. Finally, as far as plants are concerned, the target cell type must be such that cells can be regenerated into whole plants.

Techniques well known to those skilled in the art may be used to introduce nucleic acid constructs and vectors into plant cells to produce transgenic plants which comprise the heterologous fusion polypeptide coding sequence.

Agrobacterium transformation is one method widely used by those skilled in the art to transform dicotyledonous species. Production of stable, fertile transgenic plants in almost all economically relevant monocot plants is also now routine:(Toriyama, et al.

(1988) *Bio/Technology* 6, 1072-1074; Zhang, et al. (1988) *Plant Cell Rep.* 7, 379-384; Zhang, et al. (1988) *Theor Appl Genet* 76, 835-840; Shimamoto, et al. (1989) *Nature* 338, 274-276; Datta, et al. (1990) *Bio/Technology* 8, 736-740; Christou, et al. (1991) *Bio/Technology* 9, 957-962; Peng, et al. (1991) International Rice Research Institute, Manila, Philippines 563-574; Cao, et al. (1992) *Plant Cell Rep.* 11, 585-591; Li, et al. (1993) *Plant Cell Rep.* 12, 250-255; Rathore, et al. (1993) *Plant Molecular Biology* 21, 871-884; Fromm, et al. (1990) *Bio/Technology* 8, 833-839; Gordon-Kamm, et al. (1990) *Plant Cell* 2, 603-618; D'Halluin, et al. (1992) *Plant Cell* 4, 1495-1505; Walters, et al. (1992) *Plant Molecular Biology* 18, 189-200; Koziel, et al. (1993) *Biotechnology* 11, 194-200; Vasil, I. K. (1994) *Plant Molecular Biology* 25, 925-937; Weeks, et al. (1993) *Plant Physiology* 102, 1077-1084; Somers, et al. (1992) *Bio/Technology* 10, 1589-1594; WO92/14828). In particular, *Agrobacterium* mediated transformation is now a highly efficient alternative transformation method in monocots (Hiei et al. (1994) *The Plant Journal* 6, 271-282).

The generation of fertile transgenic plants has been achieved in the cereals rice, maize, wheat, oat, and barley (reviewed in Shimamoto, K. (1994) *Current Opinion in Biotechnology* 5, 158-162.; Vasil, et al. (1992) *Bio/Technology* 10, 667-674; Vain et al., 1995, *Biotechnology Advances* 13 (4): 653-671; Vasil, 1996, *Nature Biotechnology* 14 page 702). Wan and Lemaux (1994) *Plant Physiol.* 104: 37-48 describe techniques for generation of large numbers of independently transformed fertile barley plants.

Other methods, such as microprojectile or particle bombardment (US 5100792, EP-A-444882, EP-A-434616), electroporation (EP 290395, WO 8706614), microinjection (WO 92/09696, WO 94/00583, EP 331083, EP 175966, Green et al. (1987) *Plant Tissue and Cell Culture*, Academic Press) direct DNA uptake (DE 4005152, WO 9012096, US 4684611), liposome mediated DNA uptake (e.g. Freeman et al. *Plant Cell Physiol.* 29: 1353 (1984)), or the vortexing method (e.g. Kindle,

PNAS U.S.A. 87: 1228 (1990d)) may be preferred where Agrobacterium transformation is inefficient or ineffective.

Physical methods for the transformation of plant cells are reviewed in Oard, 1991, *Biotech. Adv.* 9: 1-11.

Alternatively, a combination of different techniques may be employed to enhance the efficiency of the transformation process, e.g. bombardment with Agrobacterium coated microparticles (EP-A-486234) or microprojectile bombardment to induce wounding followed by co-cultivation with Agrobacterium (EP-A-486233).

Following transformation, a plant may be regenerated, e.g. from single cells, callus tissue or leaf discs, as is standard in the art. Almost any plant can be entirely regenerated from cells, tissues and organs of the plant. Available techniques are reviewed in Vasil et al., *Cell Culture and Somatic Cell Genetics of Plants, Vol I, II and III, Laboratory Procedures and Their Applications*, Academic Press, 1984, and Weissbach and Weissbach, *Methods for Plant Molecular Biology*, Academic Press, 1989.

The particular choice of a transformation technology will be determined by its efficiency to transform certain plant species as well as the experience and preference of the person practising the invention with a particular methodology of choice. It will be apparent to the skilled person that the particular choice of a transformation system to introduce nucleic acid into plant cells is not essential to or a limitation of the invention, nor is the choice of technique for plant regeneration.

A method of making a plant cell as described herein may include introduction of a nucleic acid or a vector as described herein into a plant cell and causing or allowing recombination between the nucleic acid or vector and the plant cell genome to introduce the nucleic acid sequence into the plant cell genome.

The invention encompasses a plant cell which is transformed with a nucleic acid construct or vector as set forth above, i.e. containing a nucleic acid or vector as described above.

Within the cell, the heterologous nucleotide sequence(s) may be incorporated within the chromosome or may be extra-chromosomal. There may be more than one heterologous nucleotide sequence per haploid genome. This, for example, enables increased expression of the gene product compared with endogenous levels, as discussed below. A nucleic acid sequence comprised within a plant cell may be placed under the control of an externally inducible gene promoter, either to place expression under the control of the user or to achieve expression in response to a particular stimulus.

A plant cell may further comprise a heterologous nucleic acid sequence encoding a site-specific endoprotease, as described above. The heterologous nucleic acid sequence comprises a sequence encoding a plastid transit peptide which directs the protease to the plastid. The expressed endoprotease may be used to cleave the fusion polypeptide to liberate the recombinant polypeptide *in situ* in the plastid.

A nucleic acid which is stably incorporated into the genome of a plant is passed from generation to generation to descendants of the plant, cells of which descendants may express the encoded fusion polypeptide.

A plant cell may contain a nucleic acid sequence encoding a fusion polypeptide as described herein as a result of the introduction of the nucleic acid sequence into an ancestor cell.

In preferred embodiments, the plant cell possesses glycosylation activity which adds one or more glycan groups to the fusion polypeptide prior to localisation in the plastid.

A glycan group may be N-linked to asparagine or O-linked to serine, threonine or hydroxyproline. In preferred embodiments, the glycan is N-linked to an asparagines residue of the fusion polypeptide.

In some embodiments, the plant may possess endogenous plant glycosylation activity which adds plant specific glycans to the fusion polypeptide. Plant glycosylation involves the modification of the core $Man_3GlcNAc_2$ glycan by α1,3-fucosylation and β1, 2-xylosylation to produce a mature plant glycan which comprises α1,3 fucose and β1,2 xylose residues (Zeng et al (1997) J. Biol. Chem. 272 31340-31347).

In other embodiments, the plant may possess modified glycosylation activity which adds mammalian specific, e.g. human specific glycans to the fusion polypeptide
Mammalian glycosylation produces a mammalian glycan which comprises α1,6 fucose and does not contain xylose.

Glycosylation activity may be modified in a plant cell, for example by inhibiting endogenous plant glycosyl-transferases, such as fucosyl transferase or xylosyl transferase (Leiter H et al *J Biol Chem* (1999) 274:21830-21839) and/or expressing mammalian glycosyl-transferases, such as human 1,4 galactosyl-transferase (Lerouge, P. et al. 2000. Curr. Pharmacol. Biotechnol. 1, 347-354; Bakker, H. et al. 2001 Proc. Natl. Acad. Sci. U.S.A., 98, 2899-2904).

Methods for inhibiting gene expression and/or expressing heterologous genes in plant cells are well known in the art.

Methods described herein may further include sexually or asexually propagating or growing off-spring or a descendant of the plant regenerated from said plant cell.

A plant cell as described herein may be comprised in a plant, a plant part or a plant propagule, or an extract or derivative of a plant as described below.

Plants which include a plant cell as described herein are also provided, along with any part or propagule thereof, seed, selfed or hybrid progeny and descendants.

A plant cell may be a green algae cell, for example a Chlamydomonas spp (e.g. Chlamydomonas reinhardtii) or a Chlorella spp cell, or the plant cell may be a cell from a higher plant, for example a gymnosperm or an angiosperm. Suitable angiosperms include monocotyledons and dicotyledons.

Examples of suitable plants include tobacco, cucurbits, carrot, vegetable brassica, melons, capsicums, grape vines, lettuce, strawberry, oilseed brassica, sugar beet, Yam, wheat, barley, maize, rice, soyabeans, peas, sorghum, sunflower, tomato, potato, pepper, spinach, zinnia, chrysanthemum, carnation, poplar, eucalyptus, pine, firs and spruces.

In some preferred embodiments, cells of green algae such as Chlamydomonas or cells from dicotyledonous plants such as Arabidopsis, tobacco or poplar may be employed.

In addition to a plant, the present invention provides any clone of such a plant, seed, selfed or hybrid progeny and descendants, and any part or propagule of any of these, such as cuttings and seed, which may be used in reproduction or propagation, sexual or asexual. Also encompassed by the invention is a plant which is a sexually or asexually propagated off-spring, clone or descendant of such a plant, or any part or propagule of said plant, off-spring, clone or descendant.

A method of producing a plant may comprise incorporating nucleic acid as described above into a plant cell and regenerating a plant from said plant cell.

Another aspect of the invention provides the use of a nucleic acid, vector, cell or plant as described above in a method of producing a recombinant polypeptide as described herein.

Control experiments may be performed as appropriate in the methods described herein. The performance of suitable controls is well within the competence and ability of a skilled person in the field.

Various further aspects and embodiments of the present invention will be apparent to those skilled in the art in view of the present disclosure. All documents mentioned in this specification are incorporated herein by reference in their entirety.

Certain aspects and embodiments of the invention will now be illustrated by way of example and with reference to the figures described below.

Figure 1 shows the deduced amino acid sequence of CAH1. The arrow indicates the predicted signal peptide cleavage site. Underlined triplets indicate possible *N*-glycosylation sites.

Figure 2 shows the nucleotide sequence of Arabidopsis CAH1 mRNA.

Figure 3 shows the distribution of the antimycine A resistant NADH cytochrome *c* reductase activity and CAH1 isoforms following fractionation of the total microsome fraction from both control and BFA-treated cells over a sucrose density gradient.

Figure 4 shows the structure of the GFP-tagged and truncated forms of the Arabidopsis CAH1 protein used to localize the domain required for plastid localization. Constructs include (1-40) CAH1:GFP-fusion containing the signal peptide for the ER (first 40 amino acids), (1-103) CAH1:GFP-fusion containing the first 103 amino acids of the CAH1 and (1-40) CAH1:GFP:(224-284) CAH1 fusion containing the signal peptide for the ER (first 40 amino acids) plus the last 61 amino acid residues of the CAH1.

Experimental

Materials and Methods

Plant material and growth conditions

Arabidopsis thaliana plants, ecotype Columbia, were grown under a photon flux density of 150 µmol m^{-2} s^{-1} in a growth chamber. To obtain root material, surface-sterilized seeds (4 % sodium hypochlorite) were plated on 0.4 % agar plates supplemented with half strength Murashige and Skoog salts (Murashige, T. & Skoog, F. *Physiol. Plant.* 15, 473-497 (1962)). After three weeks, the seedlings were transferred to hydroponic conditions (Gibeaut, D.M. et al *Plant Physiol.* 115, 317-319 (1997)). The roots were sampled after two weeks.

Cloning

A putative α-CA EST clone (Arabidopsis thaliana, GenBank accession number Z18493) was used to screen a total of 3.0 x 10^5 plaques from a Uni-ZAP™ XR Arabidopsis thaliana cDNA library (Stratagene). Nucleotide sequences of three positive clones were determined and the 5´end of the cDNA was identified through 5´-RACE-PCR experiments (Gibco-BRL). A genomic library was also screened and three positive clones were subcloned. A fragment covering the 5'-end of the gene and 728 bp upstream of the putative translation initiation site was sequenced.

Southern and northern blot analysis.

Genomic DNA was extracted from developing *Arabidopsis* leaves, according to the method of Moore (Moore, D.D. Preparation of genomic DNA from plant tissue. In Current protocols in molecular biology, F.M. Ausubel et al eds (John Wiley & Sons, Inc., USA) (1994)). Total RNA was isolated from developing *Arabidopsis* leaves and roots (Verwoerd, T.C. et al *Nucl. Acids Res.* 17, 2362 (1989)). Northern blot analysis was performed as previously described (Sambrook, J. et al Molecular Cloning: A Laboratory Manual, 2nd edn. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press) (1989)).

Overexpression of recombinant CAH1 in E. coli.

PCR was used to amplify a selected cDNA region from *CAH1* and cloned into *Bam*HI -*Xho*I digested expression vector pET23a(+) (Novagen). The resulting plasmid, pSLaCAH1, verified by direct sequencing, encodes a recombinant *Arabidopsis* CAH1 starting from Gly(28), with an N-terminal T7-tag and a C-terminal 6-histidine tag. The construct was transformed into *E. coli* BL21 (DE3) and the expressed recombinant protein was purified under denaturing conditions to near-homogenity, using a histidine tag-binding resin, according to the pET System Manual (Novagen, Madison, WI, USA).

Antibody production

Polyclonal antibodies were raised against recombinant *Arabidopsis* CAH1 (Agri Sera AB, Sweden). The antibodies were purified using CAH1-coupled Affigel-10 (Bio-Rad), following the manufacturer's recommendations.

Protoplast and chloroplast isolation and fractionation.

Protoplasts were isolated from 5-10 g of *Arabidopsis* (5-7 week old) leaves, essentially according to Krömer et al (Krömer, S., et al *Plant Physiol.* 102, 947-955 (1993)), with the following slight modifications. Cell walls were digested with 1.3 % (w/v) cellulase and 0.4 % (w/v) macerase (Calbiochem) for 2 hours at 28°C without extra illumination.

Protoplasts were disrupted and chloroplasts collected as described (Kunst, L. In Methods in Molecular Biology Volume 82. *Arabidopsis* protocols, J. Martinez-Zapater and J. Salinas, eds (Totowa, NJ: Humana Press Inc.), pp. 43-53 (1998)). The chloroplasts were further purified on a 50 % (v/v) Percoll gradient (Pharmacia Biotech). The supernatant, after the disruption and centrifugation of protoplasts, represents the cytosolic fraction. This fraction was further centrifuged at 20 800 *g* at 4°C for 30 min before samples were taken for western blot and marker-enzyme assays. The residual organelle and membrane pellet was resuspended in chloroplast resuspension buffer and stored for western blot

analysis. Intact chloroplasts in chloroplast resuspension buffer were sonicated 3 x 30 s and centrifuged at 15,000 *g* for 30 min. The supernatant, mainly containing stroma proteins, was applied to a 1-mL MonoQ anion exchange column (HiTrap Q FF; Pharmacia, Sweden) equilibrated with 20 mM Tris-HCl buffer (pH 7.8). Bound proteins were eluted with a 30-mL linear gradient from 0 to 800 mM NaCl. Each fraction was desalted using PD-10 columns (Pharmacia). The purification process was monitored by subjecting aliquots from each fraction to western blot analyses.

Determination of chlorophyll and enzymatic markers.
Chlorophyll concentrations were determined in 80 % acetone according to the method of Porra et al (Porra, R.J et al *Biochim. Biophys. Acta.* 975 384-394 (1989)). The activity of the chloroplast stromal marker NADP-glyceraldehyde-3-phosphate dehydrogenase (NADP-GAPDH) was determined as described (Winter, K et al *Plant Physiol.* 69, 300-307 (1982)), phosphoenol pyruvate carboxylase (PEPc) activity was measured, as a marker for the cytosol, as described (Gardeström, P. & Edwards, G.E. *Plant Physiol.* 71, 24-29 (1983)). The activity of the ER marker NADH-cytochrome *c* reductase was determined as described (Hodges, T.K. & Leonard, R.T. *Methods Enzymol.* 32, 397-398 (1974)).

Thermolysin treatments of intact chloroplasts were performed on ice for 30 min in 40 µl reaction volumes (10 µg chlorophyll in chloroplast resuspension buffer), using 200 µg/ml thermolysin (Boehringer Mannheim).

Deglycosylation assays
A stroma fraction (100 µg protein/ml) enriched in CAH1 protein isolated from the mutant *mur1* of *Arabidopsis thaliana* was deglycosylated using a recombinant peptide-*N*-glycosidase F (PNGase F, Roche) according to the manufacturer instructions with some modifications. Samples were denatured at 100 °C for 5 min in the presence of 1% (w/v) SDS. After cooling the sample at room temperature, SDS was removed using a SDS-out kit (Pierce Co.,

Rockford, USA). The sample was then diluted with the same volume of 0.1 M Tris-HCl buffer (pH 7.8) containing 0.5 (v/v) Nonidet P-40 (Sigma). Twenty units of PNGase F were added and samples incubated for 24 and 48 h at 37°C. Samples were further analyzed by SDS-PAGE and immunoblotting with antibodies against CAH1. Fetuin (Sigma) was used as positive control during the deglycosylation experiments and treated as the stroma fractions.

2D-electrophoresis.

Stroma samples containing 300-400 µg of protein were precipitated with 0.15 % (v/v) deoxycholic acid and 72 % (v/v) TCA as described[33] and solubilized in 2D rehydration solution, containing 8 M urea, 2 % (w/v) CHAPS, and 0.002 % (w/v) bromophenol blue. The solubilized samples were loaded onto linear immobilized pH gradient gels (IPG) covering the pH ranges from 4-7 and 3-10 (Amersham Pharmacia Biotech AB, Uppsala, Sweden). The samples were applied by in-gel-rehydration and isolelectrically focused using an IPGphor system (Amersham Pharmacia Biotech AB). After focusing, strips were equilibrated twice, for 15 min each time, in equilibration buffer (50 mM Tris-HCl (pH 8.8), 6 M urea, 30 % (v/v) glycerol, 0.002 % (v/v) bromophenol blue, and 2 % (w/v) SDS), containing 1 % (w/v) DTT in the first equilibration, and 2.5 % (w/v) iodoacetamide in the second. After the equilibration steps, the strips were loaded onto 10 % SDS-PAGE gels, and electrophoretically separated at constant current. After 2D protein separation, stroma proteins were detected using a silver-staining method as described (Blum, H. et al *Electrophoresis.* 8 93-99 (1987)), or were electrotransferred onto nitrocellulose membrane. The membranes were then incubated with antibodies raised against CAH1, β(1,2)-xylose, and α(1,3)-fucose epitopes.

Mass spectrometry and protein identification.

Proteins of interest were excised from the gels and, after in-gel digestion, analyzed by mass spectrometry using a Voyager Biospectrometry Workstation (PE Biosystems, CA, USA) matrix-

assisted desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometer. The mass spectra obtained were internally calibrated using a mass standards kit (PerSeptive Biosystems, MA, USA) and used to search the NCBI database using the ProteinProspector program (available online from University of California, San Francisco). Database searches were performed using the following attributes with minor modifications, as required in each case: *Arabidopsis*, no restrictions for molecular weight and protein pI, trypsin digest, one missed cleavage allowed, cysteines modified by acrylamide, and oxidation of methionines possible, mass tolerance 50 ppm. Identification was considered positive when at least four peptides matched the protein or 30-40% coverage was obtained.

Western blot analysis.

Crude protein extracts were prepared from *Arabidopsis* leaf and root as described (Larsson, S., et al *Plant Mol. Biol.* 34, 583-592 (1997)). Protein concentration was determined using the Bio-Rad Protein Assay (Bio-Rad). SDS-PAGE was done following Laemmli (Laemmli, U. *Nature* 227, 680-685 (1970)).

Immunocytochemistry.

Developing *Arabidopsis* leaves were cut into 2 mm^2 pieces and fixed for 5 h at room temperature under a gentle vacuum. After several rinses, samples were dehydrated through a graded ethanol series and embedded in LR white resin (London Resin Co).

Immunolocalization at the light microscope level was carried out on 1-2 mm tissue sections, cut with a diamond knife on an LKB superfrost-plus microtome and then affixed to slides. The primary immune complexes were visualized by probing the sections for 2 h with colloidal gold-conjugates (6 nm) goat anti-rabbit IgG (diluted 1:100). The immuno-label was enhanced using a silver enhancement kit (Biocell), following the manufacturer's instructions, for 1 h until a black precipitate developed in the tissue. Sections were then counter-stained with toluidine blue and permanently mounted

for observation on a Zeiss Axiophot microscope using bright field illumination.

Immunolocalization at the electron microscopy level was carried out on 150 nm ultra-thin sections picked up on uncoated 200-mesh nickel grids. The gold labelling was examined on an electron microscope after staining the grids in 2% aqueous uranyl acetate for 10 min.

Expression in reticulocyte lysate in the presence of dog pancreas microsomes.

The *CAH1* gene and the N-terminally truncated version (lacking positions 1-24) were cloned into pGEM1 (Promega) with the initiator ATG codon in the context of a "Kozak consensus" sequence (Kozak, M. *Annu. Rev. Cell Biol.* 8, 197-225 (1992)). The constructs were transcribed by SP6 RNA polymerase (Promega) for 1 hour at 37°C. The transcription mixture was as follows: 1-5 µg DNA template, 5 µl 10 x SP6 H-buffer (400 mM Hepes-KOH (pH 7.4), 60 mM Mg acetate, 20 mM spermidine-HCl), 5 µl BSA (1 mg/ml), 5 µl m7G(5')ppp(5')G (10mM) (Pharmacia), 5 µl DTT (50 mM), 5 µl rNTP mix (10 mM ATP, 10 mM CTP, 10 mM UTP, 5 mM GTP), 18.5 µl H_2O, 1.5 µl RNase inhibitor (50 units), 0.5 µl SP6 RNA polymerase (20 units). Translation was performed in reticulocyte lysate in the presence or absence of dog pancreas microsomes (Hermansson, M. et al *J. Mol. Biol.* 313, 1171-1179 (2001)). The acceptor peptide Benzoyl-NLT-methylamide (Quality Control Biochemicals inc.) was added as a competitive inhibitor of glycosylation with a final concentration of 200 µM. Translation products were analyzed by SDS-PAGE and gels were quantified on a Fuji FLA-3000 phosphoimager using Fuji Image Reader 8.1j software.

Construction of GFP reporter plasmids for transient expression in Arabidopsis and tobacco cells.

The GFP reporter plasmid 35Ω-sGFP(S65T) and the plasmid containing the transit peptide (TP) sequence from RBCS fused to GFP (35Ω-TP-sGFP(S65T)) have been previously described[39]. The plasmids for expression of truncated *Arabidopsis* CAH1 protein fused to GFP were

constructed as follows: The CaMV35S-CAH1-sGFP(S65T) corresponding to the coding region of *Arabidopsis* CAH1 was PCR-amplified using the two flanking primers for-*SalI* (TAAAAGTCGACATGAAGATTATGATGATGA) and rev1-*NcoI* (AAAACCCATGGAATTGGGTTTTTCTTTTT) and the PCR product was cloned into the *SalI-NcoI* digested GFP reporter plasmid CaMV35S-sGFP(S65T). The protocol was similar for the other constructions. The CaMV35S-(1-40)CAH1-sGFP(S65T) corresponding to CAH1 containing the first 40 amino acids was PCR amplified using the two flanking primers for-*SalI* and rev2-*NcoI* (GTGTCCCATGGGGTTTGGTCCATTTTTGCC). The CaMV35S-(1-103)CAH1-sGFP(S65T) corresponding to CAH1 containing the first 103 amino acids was PCR amplified using the two flanking primers for-*SalI* and rev3-*NcoI* (TATCACCATGGCTGCTCCCTCCCCGAAGA). The CaMV35S-(1-40)CAH1-sGFP(S65T)-(224-284)CAH1 corresponding to CAH1 containing the first 40 and last 61 amino acids was PCR amplified using the two flanking primers for-*SalI* and rev2-*NcoI* and the two flanking primers for-*BsrGI* (TTCTTTGTACATCCTTGGCAAGGTGAGGTC) and rev-*BsrGI* (GACAATGTACAACTATTTTAATTGGGTTTT). The CaMV35S-CAH1-sGFP(S65T)-KDEL corresponding to the coding region of *Arabidopsis* CAH1 fused to a KDEL-tagged GFP was PCR amplified using the two flanking primers for-*SalI* and rev2-*BsrGI*:
ACAGTGTACACTAATGGTGATGGTGATGGTGATTGGGTTTTTCTTTTTGTTACC.
The plasmids were sequenced to check that the orientation and sequences of the inserted fragments were correct. The plasmids used for tissue bombardment were prepared using the QIAfilter plamid midi kit (Qiagen Laboratories).

Bombardment and fluorescence microscopy of Arabidopsis *and tobacco cells.*

Plasmids of appropriate constructions (5 µg) were introduced into *Arabidopsis* and tobacco BY2 cells using a pneumatic particle gun (PDS-1000/He; Bio-Rad). The conditions of bombardment have been previously reported (Miras, S. *et al. J. Biol. Chem.* 277, 47770-47778 (2002)). After bombardment, cells were incubated on the plates for 18-36 h (in light for the *Arabidopsis* cells, in the dark

for BY2 cells). Cells were transferred to glass slides before fluorescence microscopy.

Localization of GFP and GFP fusions was analyzed in transformed cells by fluorescence microscopy using a Zeiss Axioplan2 fluorescence microscope, and the images were captured with a digital charge-coupled devices camera, using filter sets described by Miras et al (supra).

Separation of intracellular membranes by density gradient centrifugation

Isolation of total microsome fraction and separation by density gradient centrifugation was carried out as previously described *(9)*. Briefly, ten grams of packed *Arabidopsis* cells was ground in a mortar with liquid nitrogen, resupended in 2 volumes of homogenization buffer (25 mM Tris-HCl, pH 7.5, 0.25 M sucrose, 3 mM EDTA, 1 mM DTT) and centrifuged for 15 min at 10,000 *g* at 4°C. The supernatant was centrifuged for 60 min at 150,000 *g*, supernatant (SN) was collected, an the pellet (termed total microsomes) was thoroughly resuspended in 1 mL of buffer containing 5 mM Tris-HCl, pH 7.5, 0.25 mM sucrose, 3 mM EDTA, and 1 mM DTT and loaded into a 11-mL linear gradient of 20% to 50% (w/w) sucrose buffered with 5 mM Tris-HCl, pH 7.5, 3 mM EDTA, and 1 mM DTT. Sucrose gradients were centrifuged at 80,000 *g* for 5 h at 4°C in a swing-out rotor (SW41 Beckman). Fractions (1 mL) were collected and stored at -80°C until analysis.

Brefeldin A treatment of cell suspensions

Stock solutions of brefeldin A (BFA; Sigma) were prepared at 50 mM by dissolving BFA in DMSO. Aliquots of this stock were added to 3- to 4-day-old suspension cultures to give a final concentration of 180 μM. Cells were incubated with BFA for 3 h under continuous agitation. BFA-treated cells were harvested by low-speed centrifugation.

Results

An *Arabidopsis* EST (Z18493) was identified which potentially codes for α-carbonic anhydrase (α-CA). Sequencing of the clone showed that it contained a 1046 bp open reading frame encoding a polypeptide of 284 amino acids (Figure 1). The cDNA clone was used to isolate a corresponding genomic clone, and the 5'-end of the gene and 728 bp upstream from the putative translation initiation site were sequenced. The sequence was in complete accordance with the open reading frame and upstream region of a single gene on chromosome 3 (At3g52720), which we denoted *CAH1* (U73462).

RNA was prepared from *Arabidopsis* leaf and root material and subjected to RNA blot analysis. A single hybridizing band of approximately 1200 bases was identified in leaf RNA using a fragment of the *CAH1* cDNA as a probe. No such signal was detected in root RNA. The CAH1 gene was observed to have a very pronounced diurnal variation in expression level, peaking within the first hours of the light period.

Specific antibodies raised against *Arabidopsis* CAH1 recognized a polypeptide with an apparent molecular mass of ~ 38 kDa in leaf, but not root, protein samples, confirming the northern blot data. Thus, CAH1 was observed to be expressed mainly in photosynthetic tissues.

Immunolocalization analysis was performed in *Arabidopsis* leaves to localize CAH1 within the plant cell. Unexpectedly, the results indicated that CAH1, despite its predicted sorting to the secretory pathway, was located exclusively in the chloroplast stroma.

Leaf protoplasts were fractionated into chloroplasts, cytosol and a residual organelle and membrane pellet, then assayed the CAH1 localization. Marker-enzymes for the chloroplast stroma (NADP-GAPDH) and the cytosol (PEPc) were used to assess the purity of the fractions. The activity of each enzyme in intact protoplasts was set to 100 %. A small degree of contamination (4.5 %) of

chloroplast enzymes was observed in the cytosolic fraction. The degree of contamination of the chloroplast fraction by cytosolic material was 24%, most probably due to the aggregation of chloroplasts (observed under the microscope), resulting in cytosolic enzymes being trapped. Around 60 % of the chloroplasts were intact. The broken chloroplasts explain the relatively low activity of the chloroplast marker enzyme (65 % instead of 100 %) in the chloroplast fraction. Because of the presence of a signal peptide for the ER in the unprocessed CAH1 protein, the degree of contamination of the chloroplast fraction by ER vesicles was also checked. Activity of the ER marker enzyme NADH-cytochrome *c* reductase was barely detectable in the chloroplast fraction. Nevertheless, western blot analysis, using CAH1-specific antibodies, showed that this CA is specifically located in the chloroplast fraction. A faint band was also observed in the cytosolic fraction, probably due to contamination from the broken chloroplasts. No CAH1 was found in the residual organelle and membrane pellet. The CAH1 protein in chloroplasts did not appear to be associated with the outer envelope surface, nor to protrude into the cytosol, since the protein was completely resistant to thermolysin treatment of intact chloroplasts, but susceptible after lysis of the chloroplasts. This is in accordance with the stromal localization of CAH1 observed by immunoelectron microscopy.

A translational fusion of green fluorescent protein (GFP) with the C-terminus of *Arabidopsis* CAH1 was transiently expressed in *Arabidopsis* and tobacco cells. The CAH1-GFP fusion protein was targeted to the chloroplasts in both *Arabidopsis* and tobacco cells. The expressed GFP protein (negative control) was distributed uniformly in the cytosol and in the nucleus, whereas the chloroplast control (the transit sequence of RbcS fused to GFP) was targeted to the chloroplast. Sequence information in CAH1 was therefore sufficient for chloroplast targeting of the fusion protein *in vivo*. Taken together, these findings clearly demonstrate that CAH1 is located in the chloroplast stroma of *Arabidopsis*, despite the presence of a typical ER-targeting signal peptide.

For further examination of the domain required for chloroplast localization of the CAH1 protein, several versions of the CAH1 protein were generated and the effects of transiently expressing corresponding GFP fusions in *Arabidopsis* and BY2 tobacco cells were tested. The first 40 amino acid residues of CAH1, containing the predicted ER signal peptide, were fused to GFP containing an ER retention signal (KDEL) in the C-terminus. This fusion protein was found to be retained in the ER, showing that the CAH1 ER signal peptide is functional and sufficient for targeting the protein to the secretory pathway. In addition, when the full-length protein was fused to GFP containing an ER retention signal (KDEL), the fusion was also retained in the ER, thus ruling out that any domain in the mature protein blocks ER targeting. No GFP activity was observed in the chloroplasts for any of the constructs tested.

In vitro uptake studies were performed both with isolated chloroplasts, and with ER-derived dog pancreas microsomes (Monné, M. et al *J. Biol. Mol.* 293, 807 (1999)). Intact pea chloroplasts were not able to take up or process the CAH1 precursor, providing indication that the translocation of CAH1 across the envelope membranes may not take place through the Tic/Toc translocon system. Efficient uptake, signal peptide processing, and glycosylation were observed with microsomes. The ER signal peptide is required for uptake of the protein into the microsomes, since a truncated CAH1 form, lacking this signal is not taken up into the ER, as evidenced by lack of glycosylation and sensitivity to externally added proteinase K. With full-length CAH1, the signal peptide is cleaved off after import into the microsomes and this process leads to a small shift in mobility.

The CAH1 protein has five predicted acceptor sites for *N*-linked glycosylation (Fig. 1), and major products with relative molecular masses of approximately 38, 41 and 44 kDa were observed in addition to the non-modified 31-kDa protein. The addition of a competitive glycosylation peptide inhibitor prevents the occurrence of the high molecular weight products, providing indication that at least four

glycosylation sites may be partially modified. Removal of the signal peptide leads only to a small shift in mobility and a product corresponding to the protein lacking the signal peptide is clearly seen when glycosylation is blocked. The glycosylated forms and the unglycosylated, signal-peptidase cleaved forms of the protein are resistant to externally added proteinase K and are located in the lumen of the microsomes. These findings provide indication that CAH1 is taken up by the ER and glycosylated before being targeted into the chloroplast.

Brefeldin A (BFA) is a fungal antibiotic that inhibits Golgi-mediated vesicular traffic (C. Ritzenthaler, *et al. Plant Cell* 14, 237 (2002)). The effect of BFA on the intracellular distribution of CAH1 was analysed in different sub-cellular fractions isolated from *Arabidopsis* cell suspensions. *Arabidopsis* cells were treated for 3 h in the absence (control) and presence of 180 μM BFA. Supernatant (SN) and total microsome fraction (MS) were obtained as described in Materials and Methods. All the fractions were immunoblotted with antibodies against CAH1 with five μg proteins loaded in each lane. Antimycine A resistant NADH cytochrome *c* reductase activity (nmol NADH mg $prot^{-1}$ min^{-1}) was also measured in the supernatant and in the total microsome fractions.

In the absence of BFA, the mature CAH1 form was observed to accumulate in the soluble fraction. Under these conditions, a minor low molecular mass form corresponding to the unglycosylated CAH1 precursor was found in the microsomal fraction. In the presence of BFA, accumulation of the mature CAH1 form in the soluble fraction was found to be greatly reduced. However, BFA caused strong accumulation of both CAH1 precursor and partially glycosylated CAH1 forms in the microsomal fraction.

Further separation of fractions from both control and BFA treated cells by sucrose density gradients showed that these CAH1 forms were localized in light dense microsomes, particularly in ER-rich

fractions (Fig. 3). This indicates that vesicular transport along the Golgi apparatus is an intermediate step in the trafficking of CAH1 to the chloroplast.

Despite its chloroplast localization, CAH1 has an N-terminal signal peptide that targets the protein to the ER. Stroma were isolated from *Arabidopsis* chloroplasts and fractionated it by anion exchange chromatography. The CAH1-containing fraction was separated by 2D-gel electrophoresis, and either silver stained or blotted onto nitrocellulose membranes. The membranes were then incubated with antibodies raised against CAH1, $\beta(1,2)$-xylose, and $\alpha(1,3)$-fucose epitopes. These two antibodies recognize xylose- and fucose-containing glycans *N*-linked to Asn-x-Thr/Ser sites, respectively (Faye, L. et al. *Anal. Biochem.* 209, 104-108 (1993)): linkages that are typical of plants and are specifically transferred to *N*-glycans within the Golgi apparatus (Lerouge, P., *et al. Plant Mol. Biol.* 38, 31-48 (1998)). Antibodies raised against CAH1 cross-reacted with a protein at ~38 kDa with a variable pI value ranging from 5.2 to 5.6 (Fig. 5b). Antibodies raised against $\beta(1,2)$-xylose and $\alpha(1,3)$-fucose cross-reacted with the same protein recognized by the CAH1 antibodies, providing indication that the mature stromal CAH1 protein is *N*-glycosylated.

CAH1 was not the only glycosylated protein found to be present in the stroma of *Arabidopsis*. By comparing 2D- gels (covering the pH ranges from 4-7 and 3-10) from different stroma preparations, we have identified approximately 6-10 different spots that cross-react with both xylose and fucose antibodies.

Some of these protein spots were excised and subjected to MALDI-TOF MS analysis, which positively identified a putative chloroplast 50S ribosomal protein (At1g05190.1; spot no. 1) and an unknown protein (At4g04240.1; spot no. 2). NetNGlyc analysis for predicting potential *N*-glycosylation sites (Gupta R & Brunak S (2002) Pac. Symp. Biocomput. 310-322) strongly predicts that 1-3 acceptor sites

for *N*-linked glycosylation are contained in the sequence of these two proteins. These data show that *N*-glysosylation of stromal proteins in *Arabidopsis thaliana* is not restricted to CAH1.

The C-termini of both CAH1 and the putative chloroplast 50S ribosomal protein show high degrees of similarity. They are extremely hydrophilic (16 of 19 residues, and nine of the last 15 C-terminal amino acid residues, are charged, including six and five lysine residues, respectively). This C-terminus may be important for the mechanism whereby these proteins are imported to the chloroplast.

The data herein provides firm evidence that the chloroplast proteome contains glycosylated proteins which are sorted through the ER, in addition to those proteins which are synthesized in the chloroplast and those which are transported through the Tic/Toc translocon complex.

Since different types of plastid are of similar origin and can re-develop into each other, these findings have significant application in the expression of recombinant plastid polypeptides.

Gene ID	Description	NA Acc No:	AA Acc no
AT1G03860	prohibitin 2 -related B-cell receptor associated protein	NM_202027	NP_973756
AT1G09180	GTP-binding protein SAR1, putative strong similarity to SP:Q01474 GTP-binding protein SAR1B and SP:O04834 GTP-binding protein SAR1A [Arabidopsis thaliana]	NM_100788	NP_172390
AT1G13900	calcineurin-like phosphoesterase family contains Pfam profile: PF00149 calcineurin-like phosphoesterase	NM_101256	NP_172843
AT1G15690	inorganic pyrophosphatase -related similar to inorganic pyrophosphatase GI:790478 from [Nicotiana tabacum]	NM_101437	NP_173021
AT1G26560	glycosyl hydrolase family 1 similar to beta-glucosidase GB:L41869 GI:804655 from [Hordeum vulgare]	NM_102418	NP_173978
AT1G29670	"GDSL-motif lipase/hydrolase protein similar to family II lipase EXL1 GI:15054382 from [Arabidopsis thaliana]; contains Pfam profile: PF00657 Lipase/Acylhydrolase with GDSL-like motif"	NM_102707	NP_174260
AT1G30360	ERD4 protein nearly identical to ERD4 protein (early-responsive to dehydration stress) [Arabidopsis thaliana] GI:15375406; contains Pfam profile PF02714: Domain of unknown function DUF221	NM_102773	NP_564354
AT1G33590	"disease resistance protein-related (LRR) contains leucine rich-repeat domains Pfam:PF00560, INTERPRO:IPR001611; similar to Hcr2-5D [Lycopersicon esculentum] gi|3894393|gb|AAC78596"	NM_103082	NP_564426
AT1G47128	cysteine proteinase RD21A identical to thiol protease RD21A SP:P43297 from [Arabidopsis thaliana]	NM_103612	NP_564497
AT1G49750	leucine rich repeat protein family contains leucine-rich repeats, Pfam:PF00560	NM_103862	NP_175397
AT1G61790	Hypothetical protein	NM_104861	NP_176372
AT1G66770	"nodulin MtN3 family protein contains Pfam PF03083 MtN3/saliva family; similar to LIM7 (cDNAs induced in meiotic prophase in lily microsporocytes) GI:431154 from [Lilium longiflorum]"	NM_105348	NP_176849
AT1G68560	glycosyl hydrolase family 31 (alpha-xylosidase) identical to alpha-xylosidase precursor GB:AAD05539 GI:4163997 from [Arabidopsis thaliana]	NM_105527	NP_177023
AT1G74180	"leucine rich repeat protein family contains	NM_106078	NP_177558

	leucine rich-repeat (LRR) domains Pfam:PF00560, INTERPRO:IPR001611; similar to Hcr2-0B [Lycopersicon esculentum] gi|3894387|gb|AAC78593"		
AT2G06850	xyloglucan endotransglycosylase (ext/EXGT-A1) identical to endo-xyloglucan transferase (ext) GI:469484 and endoxyloglucan transferase (EXGT-A1) GI:5533309 from [Arabidopsis thaliana]	NM_126666	NP_178708
AT2G10940	"protease inhibitor/seed storage/lipid transfer protein (LTP) family similar to proline-rich cell wall protein [Medicago sativa] GI:3818416; contains Pfam profile PF00234 Protease inhibitor/seed storage/LTP family"	NM_179618	NP_849949
AT2G22170	expressed protein	NM_127785	NP_565527
AT2G37290	Hypothetical protein and genefinder	NM_129285	NP_181266
AT2G45740	expressed protein	NM_180110	NP_850441
AT3G05660	"disease resistance protein family contains leucine rich-repeat (LRR) domains Pfam:PF00560, INTERPRO:IPR001611; similar to Cf-2.2 [Lycopersicon pimpinellifolium] gi|1184077|gb|AAC15780"	NM_111439	NP_187217
AT3G14210	"myrosinase-associated protein, putative similar to GB:CAA71238 from [Brassica napus]; contains Pfam profile:PF00657 Lipase/Acylhydrolase with GDSL-like motif"	NM_112278	NP_188037
AT3G14590	"C2 domain-containing protein low similarity to SP|Q16974 Calcium-dependent protein kinase C (EC 2.7.1.-) (Aplysia californica); contains Pfam profile PF00168: C2 domain"	NM_112319	NP_188077
AT3G16240	delta tonoplast integral protein (delta-TIP) identical to delta tonoplast integral protein (delta-TIP) GB:U39485 [Arabidopsis thaliana] (Plant Cell 8 (4), 587-599 (1996))	NM_112495	NP_188245
AT3G20820	"disease resistance protein family (LRR) contains similarity to Cf-2.1 [Lycopersicon pimpinellifolium] gi|1184075|gb|AAC15779; contains leucine rich-repeat domains Pfam:PF00560, INTERPRO:IPR001611"	NM_112973	NP_188718
AT3G27280	prohibitin -related similar to prohibitin GB:AAC49691 from [Arabidopsis thaliana] (Plant Mol. Biol. (1997) 33 (4), 753-756)	NM_202640	NP_974369
AT3G54110	uncoupling protein (ucp/PUMP)	NM_115271	NP_190979
AT3G54400	nucleoid DNA-binding - like protein nucleoid DNA-binding protein cnd41, chloroplast, common tobacco, PIR:T01996	NM_115300	NP_191008
AT3G55200	"splicing factor, putative contains CPSF A subunit region (PF03178); contains weak WD-40 repeat (PF00400); similar to Splicing factor 3B subunit 3 (SF3b130)/spliceosomal protein/Splicing	NM_115378	NP_567015

	factor 3B subunit 3 (SAP 130)(KIAA0017)(SP:Q15393) Homo sapiens, EMB		
AT4G17340	major intrinsic protein (MIP) family contains Pfam profile: MIP PF00230	NM_117838	NP_193465
AT4G27520	expressed protein ENOD20 gene, Medicago truncatula, X99467	NM_118887	NP_194482
AT4G39730	expressed protein	NM_120134	NP_195683
AT5G02260	"expansin, putative (EXP9) similar to expansin precursor GI:4138914 from [Lycopersicon esculentum]; alpha-expansin gene family, PMID:11641069"	NM_120304	NP_195846
AT5G03350	expressed protein	NM_120414	NP_195955
AT5G07340	"calnexin, putative identical to calnexin homolog 2 from Arabidopsis thaliana [SP|Q38798], strong similarity to calnexin homolog 1, Arabidopsis thaliana, EMBL:AT08315 [SP|P29402]; contains Pfam profile PF00262 calreticulin family"	NM_120816	NP_196351
AT5G12860	Oxoglutarate/malate translocator, putative similar to 2-oxoglutarate/malate translocator precursor, spinach, SWISSPROT:Q41364	NM_121289	NP_568283
AT5G25980	glycosyl hydrolase family 1 similar to myrosinase precursor (EC 3.2.3.1)(Sinigrinase) (Thioglucosidase) SP|P37702 from [Arabidopsis thaliana]	NM_122499	NP_568479
AT5G26000	glycosyl hydrolase family 1, myrosinase precursor	NM_122501	NP_197972
AT5G26260	expressed protein various predicted proteins, Arabidopsis thaliana	NM_122527	NP_568483
AT5G44020	vegetative storage protein-related	NM_123769	NP_199215
AT5G63840	glycosyl hydrolase family 31 similar to alpha-glucosidase GI:2648032 from [Solanum tuberosum]	NM_125779	NP_201189
AT5G65760	"hydrolase, alpha/beta fold family similar to SP|P42785 Lysosomal Pro-X carboxypeptidase precursor (EC 3.4.16.2) (Prolylcarboxypeptidase) (PRCP) (Proline carboxypeptidase) (Homo sapiens); contains Pfam profile PF00561: hydrolase, alpha/beta fold family"	NM_125973	NP_201377
At2g31910	putative Na+/H+ antiporter	NM_128749	NP_180750
At2g01720	Ribophorin I-like protein	NM_126233	NP_178281
At4g20990	Carbonic anhydrase	NM_118217	NP_193831
At4g39730	Expressed protein	NM_120134	NP_195683
At1g21750	Protein disulfide isomerase	NM_179365	NP_849696

Table 1